Traitement automatique du bambara
Objectifs et premiers résultats

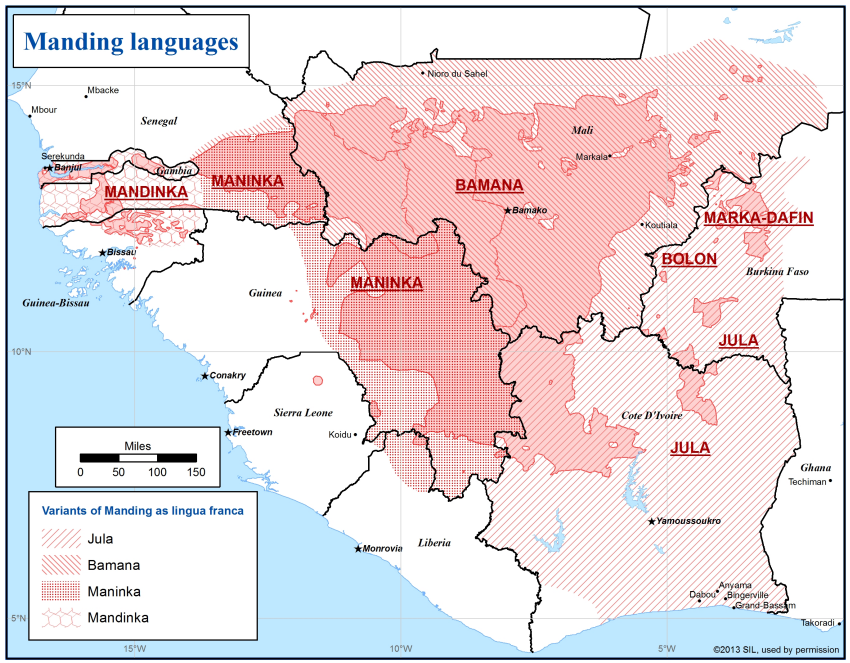
Damien Nouvel
Valentin Vydrin
Davy Auffret

ERTIM
Equipe de recherche
textes, informatique,
multilinguisme

Langage, Langues et
Llacan
Cultures d'Afrique Noire

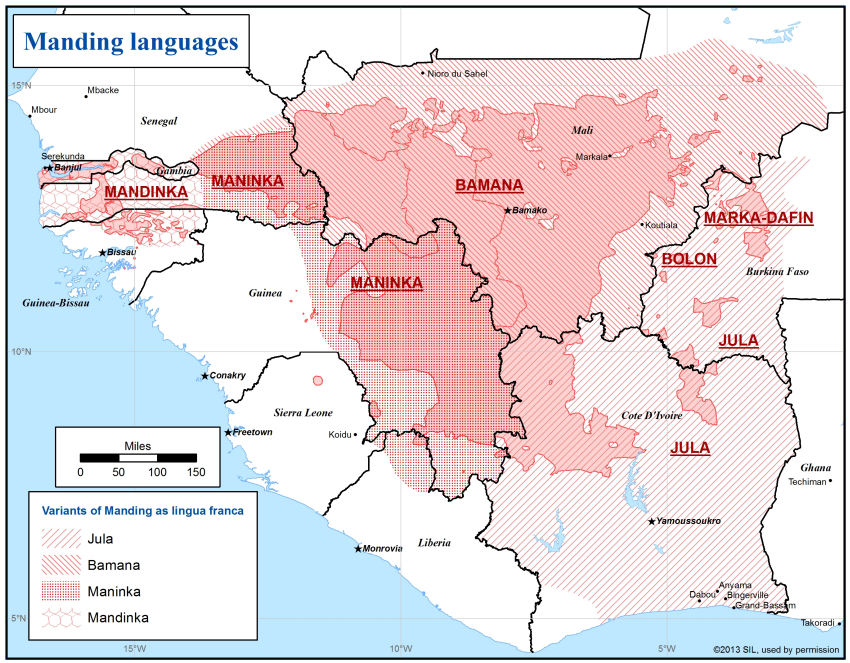
Plan

Contexte



- Parlée principalement au Mali (diglossie : français, 20%)
- Ou « bamanaka »

Contexte



- Parlée principalement au Mali (diglossie : français, 20%)
- Ou « bamanaka »
- 4M de locuteurs (10M en 2 ème langue)
- Véhiculaire, tradition orale
- Macro-langue mandingue (avec dioula, malinké, etc.)

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

$\Rightarrow$ Langues peu standardisées

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

$\Rightarrow$ Langues peu standardisées

- Grammaire
  - Type : S AUX O V X, tonale
  - Pas de genre grammatical
  - Pas de conjugaison (marques prédicatives AUX)
  - Peu de flexion (-w : pluriel)

Travaux antérieurs

- Références sur le bambara
  - (Binger, 1886) Essai sur la langue bambara.
  - (Sauvant, 1926) Dictionnaire bambara-français.
  - (Vydrin, 1999) Les parties du discours en bambara.
  - (Vydrin, 1999) Manding-english dictionary (maninka, bamana).
  - (Dumestre, 2003) Grammaire fondamentale du bambara.
  - (Bailleul, 2007) Dictionnaire bambara-français.
  - (Vydrin, 2008) Glossed electronic corpora of Mande languages.
  - (Dumestre, 2011) Dictionnaire bambara-français.
  - (Vydrin, 2011) Corpus électronique annoté des textes bambara.
  - (Enguehard, 2012) Vers l'info. de langues d'Afrique de l'Ouest.
  - (Maslinsky, 2014) Daba : a model and tools for manding corpora.

Plan

Corpus bambara de référence

- Collecte de textes en bambara
  - Publiés (périodiques, littérature) ou non (lettres, trans.)
  - Normalisation des textes (orthographe, tons, etc.) ?
  - Textes en ligne `http://cormand.huma-num.fr/biblio/`

$\Rightarrow$ Volume : 2,3M mots

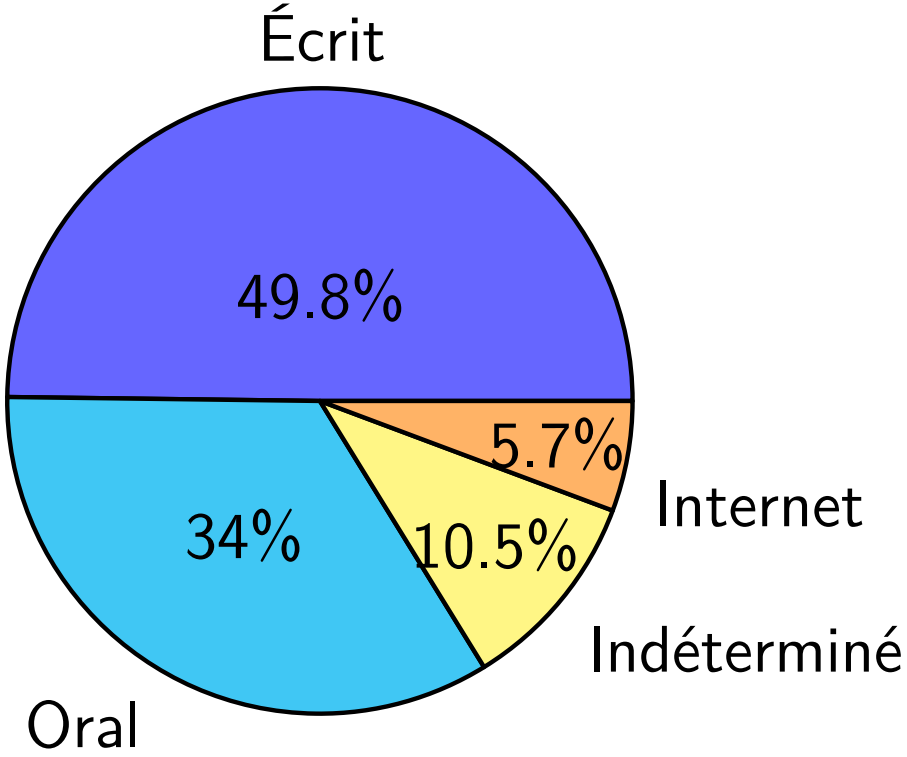
- Utilisation essentiellement linguistique
  - Apprentissage de la langue
  - Études linguistiques sur corpus
  - Annotation (parties du discours, lemmes, morphologie, gloses)

$\Rightarrow$ Labex EFL, axe 6 : ressources linguistiques

$\Rightarrow$ Site internet : `http://cormand.huma-num.fr/`

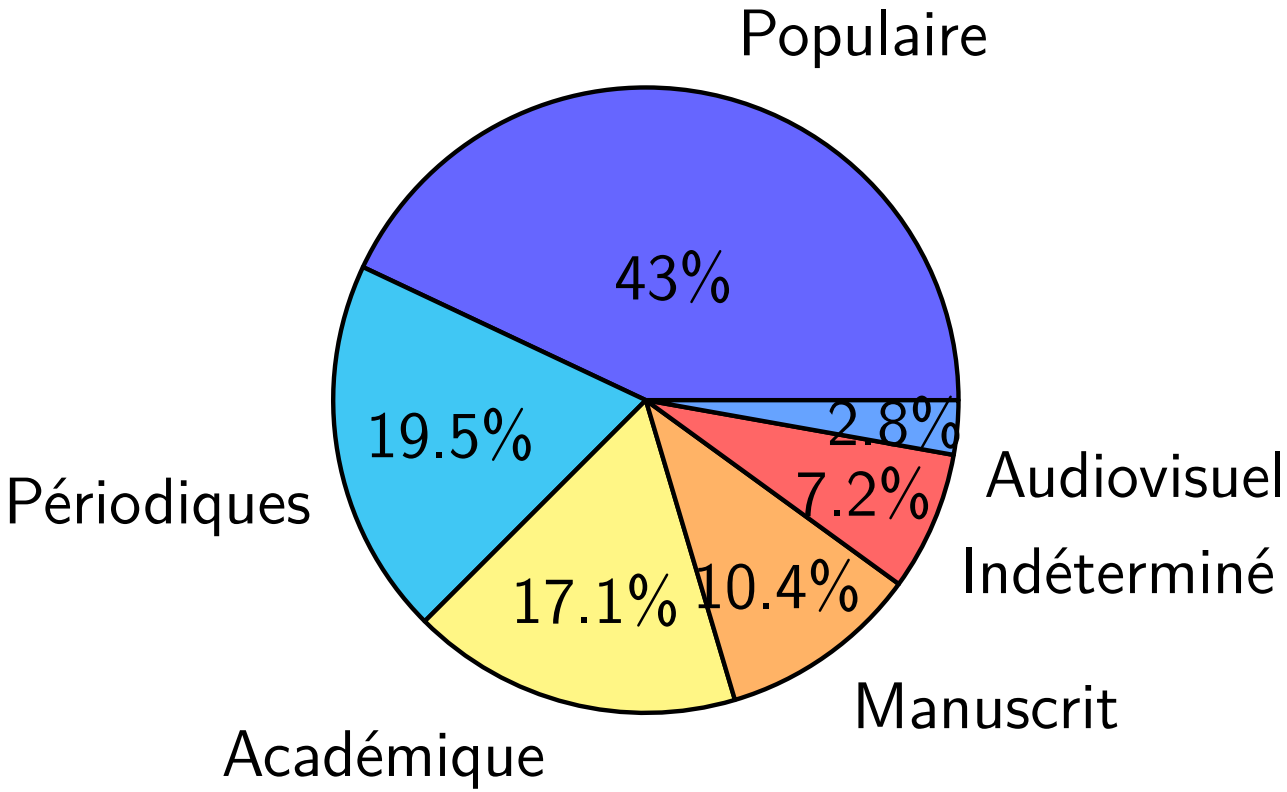
$\Rightarrow$ Modélisation linguistique informatisée (SketchEngine)

Mediums du corpus



$\Rightarrow$ Prédominance de l'écrit, mais souvent issu de l'oral (contes, etc.)

Sources du corpus



Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

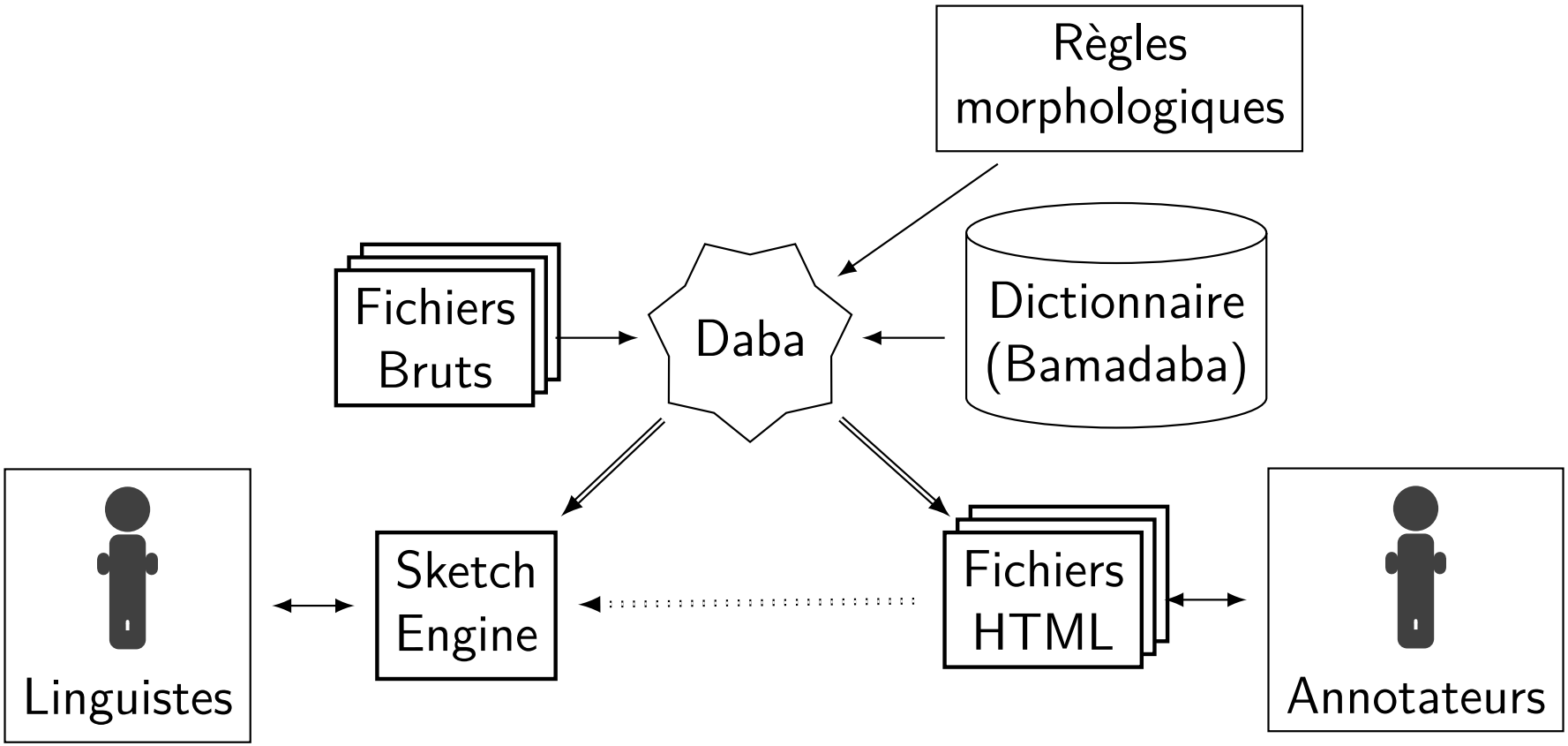
Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

- Annotation manuelle
  - Peu de moyens (profs, étudiants, bénévoles)
  - Validation humaine
  - Correction / normalisation

Schéma de fonctionnement



Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - (Entités nommées)
  - (Malinké)

$\Rightarrow$ Exploitation des données annotées (supervision)

Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - (Entités nommées)
  - (Malinké)

$\Rightarrow$ Exploitation des données annotées (supervision)

- Cadre général
  - Collaboration LLACAN / ERTIM
  - Durée : 3 ans
  - Budget interne Inalco (stages, prestations, missions)

$\Rightarrow$ Interactions avec les linguistes

Plan

Procédures d'annotation

- Plusieurs niveaux de traitement

Procédures d'annotation

- Plusieurs niveaux de traitement
  - Fichier source

Procédures d'annotation

- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres

Procédures d'annotation

- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique

    $\Rightarrow$ Automatique, sortie ambiguë

Procédures d'annotation

- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique

    $\Rightarrow$ Automatique, sortie ambiguë
  - Annotation par les linguistes

$\Rightarrow$ Ambiguïtés résiduelles

Procédures d'annotation

- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique

    $\Rightarrow$ Automatique, sortie ambiguë
  - Annotation par les linguistes

  $\Rightarrow$ Ambiguïtés résiduelles

$\Rightarrow$ Plusieurs versions de fichiers à synchroniser

Fichier source

<h>Diɲɛ Yaalala</h>
Nsiirin, nsiirin. N y'a bila den dɔ le kan.
Den nin ye sira dali a facɛ fɛ, k'a b'a fɛ ka taga diɲɛ yaala ka dɔ fara a hakili kan. A facɛ ye sira d'a ma a ka taga yaala. A tagara yaala kosɛbɛ.
A fɔlɔla ka taga ben sogosu dɔ ma. A yɔrɔ bɛɛ tolira, a ko : "E ! Ala bɛ se." Sogo nin wulila ka kum'a fɛ k'a kan'a fɔ ko Ala bɛ se, k'Ala ka se b'a ɲɛfɛ. "A tɛmɛna sogo nin na ka taga ɲɛfɛ, ka kɔlɔn saba ye. Fɔlɔ jalen bɛ, ji foyi t'a la. A filanan ji to kɔnɔ. A sabanan ji b'o kɔnɔ. A tɛmɛn'o la ka taga se cɛkɔrɔnin dɔ ma. Cɛ nin kɔrɔla kosɛbɛ. [...]

Fichier Daba

Nsiirin, nsiirin.

Nsiirin	,	nsiirin	.
nsíirin		nsíirin	
n		n	
conte		conte	

N y'a bila den dɔ le kan.

N	y'	a	bila	den	dɔ	le	kan	.
ń	y'	à	bìla	dén	dɔ́	le	kàn	
pers	pm	pers	v	n	dtm	prt	pp	
1SG	PFV.TR	3SG	mettre	enfant	certain	FOC	sur	

Fichier vertical

Token	Lemma	PdD	Glose	Compos.	Original	Tonal
nsiiri	nsiiri	n	conte		NSIIRI	nsiiri
naaninan	naaninan	ORD|adj	quatrième	naani	NAANINAN	náaninan
dinye	dunya|jinye |dununye|dinye |jyen|diyen	n	monde		Diɲɛ	díɲɛ
yaalala	yaalala	AG.PRM|n		yaala	Yaalala	yáalala
nsiirin	nsiiri|nsiirin	n	conte		Nsiirin	nsíirin
,	,	c	,	,	,	,
nsiirin	nsiiri|nsiirin	n	conte		nsiirin	nsíirin
.	.	c	.	.	.	.
n	n	pers	1SG		N	ń
y'	ye|y'	pm	PFV.TR		y'	y'
a	a	pers	3SG		a	à
bila	bil'|bla|bila	v	mettre		bila	bìla
den	den	n	enfant		den	dén
do	do	dtm	certain		dɔ	dɔ́
le	le	prt	FOC		le	le
kan	kan	pp	sur		kan	kàn
.	.	c	.	.	.	.

Volumétrie des ressources

- Corpus

Corpus	**Balises**	**Ponctuations**	**Formes (distinctes)**
Brut	412K	383K	2 321K (68K)
Désambiguïsé	104K	71K	426K (19K)

Volumétrie des ressources

- Corpus

Corpus	Balises	Ponctuations	Formes (distinctes)
Brut	412K	383K	2 321K (68K)
Désambiguïsé	104K	71K	426K (19K)

- Dictionnaires (disponibles en ligne)

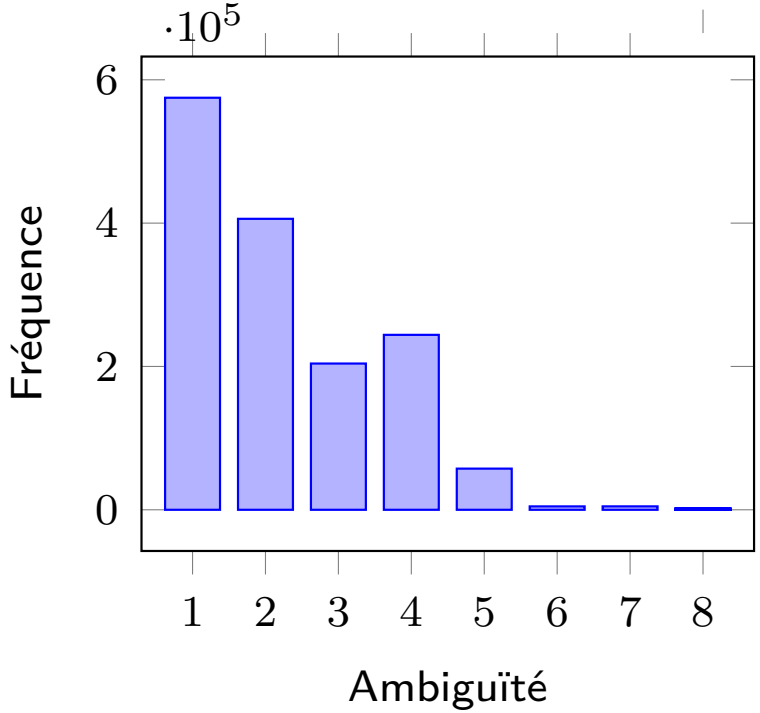
Dictionnaire	Description	Entrées	Ambiguïté
bamadaba	Dictionnaire principal	11K	1,137
enciclop	Notions encyclopédiques	29	1
jamuw	Noms claniques	375	1,001
togow	Prénoms	496	1
yorow	Toponymes	299	1

Statistiques sur les ambiguïtés

- Corpus échantillon (1,5M mots)
  - Traité avec Daba
  - 1,285M mots (215K ponctuations)
  - 72K (5,6%) parties du discours non renseignées
  - 288K (22%) non-ambiguës (selon Daba)

Statistiques sur les ambiguïtés

- Corpus échantillon (1,5M mots)
  - Traité avec Daba
  - 1,285M mots (215K ponctuations)
  - 72K (5,6%) parties du discours non renseignées
  - 288K (22%) non-ambiguës (selon Daba)



Ambiguïté	Fréquence
pers|intj	75K
n|v	72K
cop|pp|v|pm	57K
adv|conj|pm|v	53K
cop|pm	43K
v|n	38K
prt|conj|prn	33K
n.prop|pp|conj|intj	23K
conj|n|prep|v	23K
vq|v|adj|n|cop	22K

Caractérisation des ambiguïtés

- Ambiguïtés fréquentes

Mot	Ambiguïté	Fréquence
à	pers|intj	74K
yé	cop|pp|v|pm	58K
kà	adv|conj|pm|v	53K
ò	prt|conj|prn	34K

Caractérisation des ambiguïtés

- Ambiguïtés fréquentes

Mot	Ambiguïté	Fréquence
à	pers|intj	74K
yé	cop|pp|v|pm	58K
kà	adv|conj|pm|v	53K
ò	prt|conj|prn	34K

- Par mots distincts

Ambiguïté	Fréquence
PL|n	1304
n|v	1189
v|n	1039
ptcp|PTCP.RES	588
PFV.INTR|v	584
NMLZ|n	568
n.prop|dtm|ptcp|prn|adj|PL|n	497
n.prop|n	483

Jeu d'étiquettes sur le corpus désambiguisé

Code	Partie du discours	Quantité
n	nom	82K
c	ponctuation	66K
pers	pronom personnel	54K
v	verbe	51K
pm	marque prédicative	41K
pp	postposition	34K
conj	conjonction	21K
cop	copule	18K
n.prop	nom propre	12K
dtm	déterminatif	12K
prn	pronom (non-personnel)	10K
prt	particule	10K

num	numératif	6K
adj	adjectif	4K
ptcp	participe	4K
intj	interjection	2K
adv	adverbe	2K
vq	verbe qualitatif	1K
onomat	onomatopée	102
adv.p	adverbe préverbal	26
conv.n	converbe nu	24
mrph	morphème	13

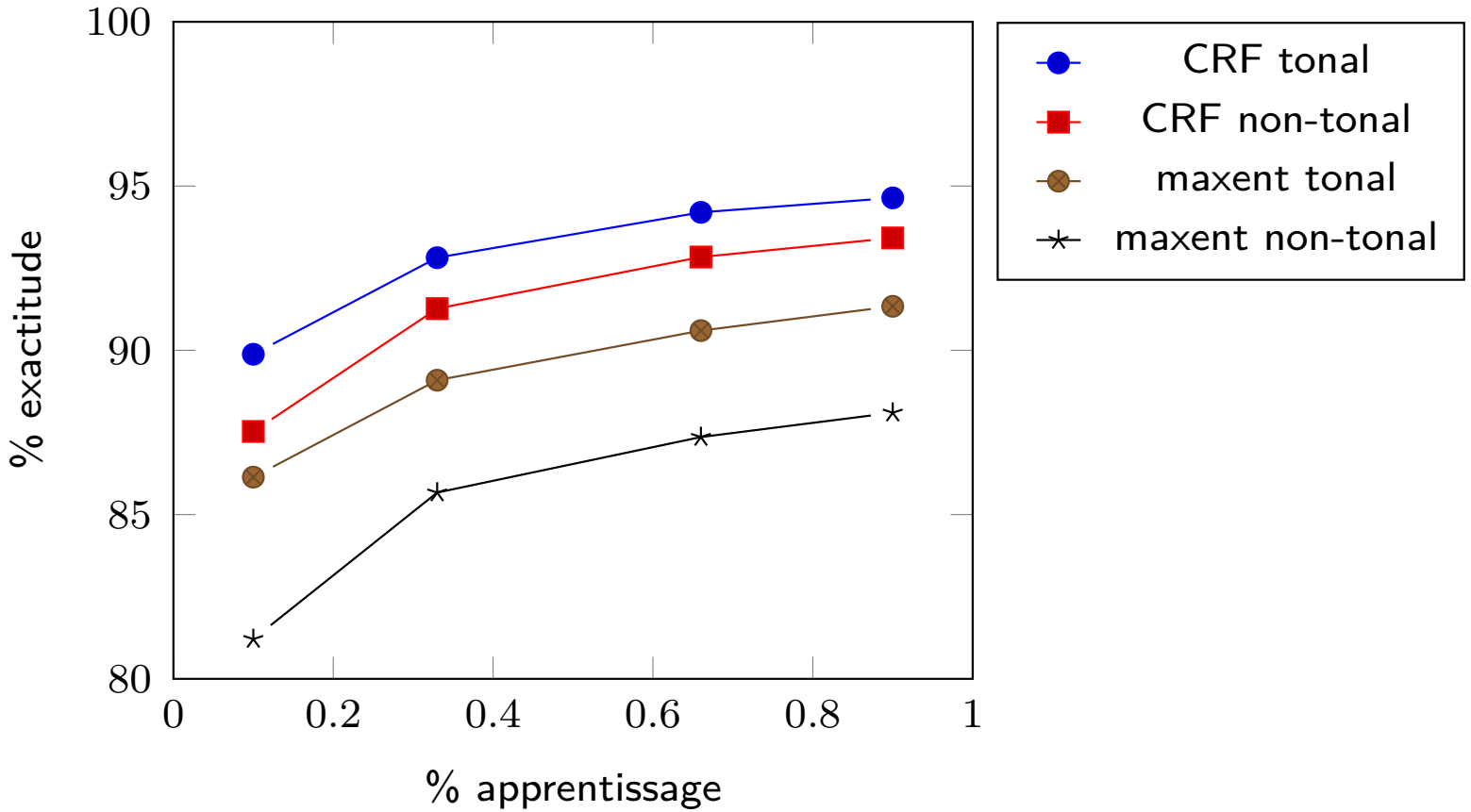
$\Rightarrow$ Jeu d'étiquettes relativement standard

Apprentissage : algorithme et données

- Méthodologie
  - Utilisation de Wapiti (Lavergne, 2010)
  - Validation croisée à 10 plis

Apprentissage : algorithme et données

- Méthodologie
  - Utilisation de Wapiti (Lavergne, 2010)
  - Validation croisée à 10 plis



Médiums avec corpus complets

- Méthodologie
  - Sélection une sous-partie pour l'apprentissage
  - Évaluation une autre sous-partie (sans validation croisée)

Médiums avec corpus complets

- Méthodologie
  - Sélection une sous-partie pour l'apprentissage
  - Évaluation une autre sous-partie (sans validation croisée)

App \ Test	**Écrit**	**Internet**	**Oral**	**Indet.**
Écrit	*98.55*	92.0	90.36	**92.93**
Internet	83.78	*99.08*	82.63	**84.20**
Oral	88.43	88.92	*98.81*	**90.68**
Indet.	87.26	**88.26**	87.07	*99.21*

Médiums avec corpus équilibrés

- Méthodologie
  - Sélection une sous-partie pour l'apprentissage
  - Sous-échantillonnage selon corpus d'évaluation : 25K éléments à la phrase près
  - Évaluation une autre sous-partie

Médiums avec corpus équilibrés

- Méthodologie
  - Sélection une sous-partie pour l'apprentissage
  - Sous-échantillonnage selon corpus d'évaluation : 25K éléments à la phrase près
  - Évaluation une autre sous-partie

App \ Test	Écrit	Internet	Oral	Indet.
Écrit	*98.97*	**86.09**	85.79	**86.09**
Internet	85.52	*99.08*	84.22	82.68
Oral	85.56	85.24	*99.08*	85.78
Indet.	**85.92**	84.70	**86.27**	*99.36*

Traits pour l'apprentissage

- Ajouts de traits pour
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Une version transformé du mot par un expression régulière, ANP
  - Une version non-tonalisée du mot
  - La taille du mot
  - Les étiquettes possibles dans les dictionnaires

Traits pour l'apprentissage

- Ajouts de traits pour
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Une version transformé du mot par un expression régulière, ANP
  - Une version non-tonalisée du mot
  - La taille du mot
  - Les étiquettes possibles dans les dictionnaires

Score	Base	Non-tonal	Préf.	Suf.	Dico	ANP	Taille	Tous
Mot U	86.14	86.36	89.09	89.46	89.87	89.41	89.14	90.65
Phrase U	18.70	19.09	24.38	25.60	25.62	25.18	23.82	29.04
Mot B	85.94	91.27	91.69	91.91	91.50	86.84	88.37	94.22
Phrase B	13.14	35.35	36.39	37.51	33.94	15.20	21.57	47.90

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Outil	**Score**
Baseline	22%
Majorité	82,06%
TreeTagger	93,50
Wapiti	94.22
Majorité (non-tonal)	79,68%

Plan

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)

  $\Rightarrow$ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)

  $\Rightarrow$ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

$\Rightarrow$ La présence de tons aide pour la morpho-syntaxe

$\Rightarrow$ Détecter automatiquement les tons ?

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

$\Rightarrow$ 1 518 tonalisations ambiguës, en moyenne 2,26 tonalisations

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

$\Rightarrow$ 1 518 tonalisations ambiguës, en moyenne 2,26 tonalisations

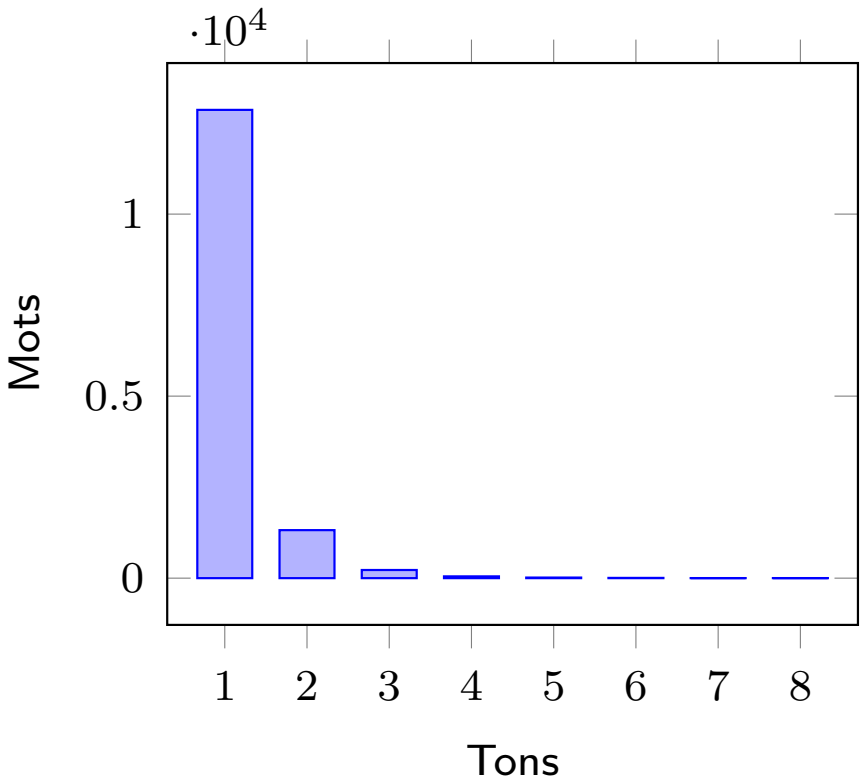
$\Rightarrow$ Baseline, tonalisation la plus fréquente : 59% (dont ponctuations)

Répartitions des tons

- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots

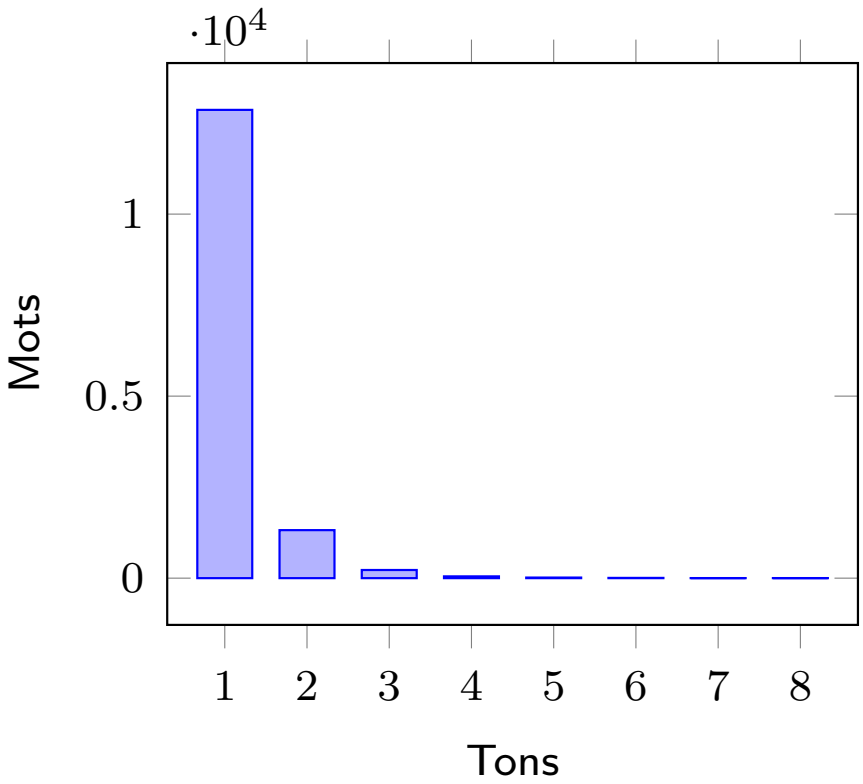
Répartitions des tons

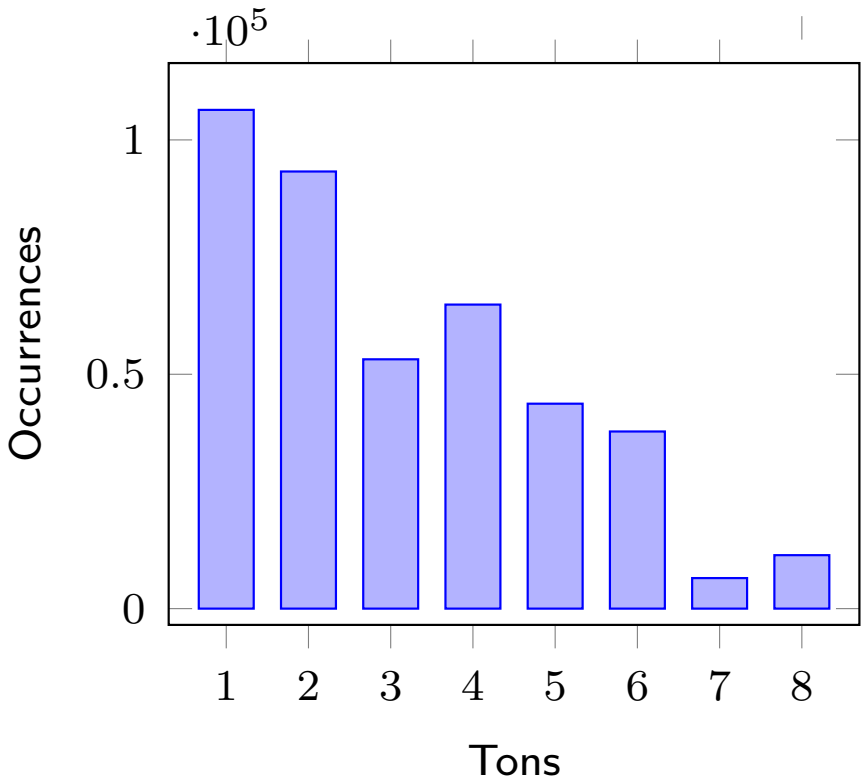
- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots



Répartitions des tons

- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots





Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Ent.	Tonalisations	Traduction
3.20	táatúmà : 1.0 , tàatúmá : 1.0 , táatúmá : 1.0 , táatùma : 9.0 , táatúma : 1.0 , tàatúmà : 1.0 , tàatúma : 1.0 , tàatùmà : 1.0 , tàatùmá : 1.0 , táatùmá : 1.0 , táatùmà : 2.0 , tàatuma : 1.0 , tàatùma : 1.0 , táatuma : 1.0	Départ (?)
2.50	bámànankɛ : 10.0 , bámànakɛ̀ : 16.0 , bámànankɛ̀ : 5.0 , bámànánkɛ : 6.0 , cɛ̀ : 16.0 , bámànanke : 1.0 , bámànan : 16.0 , bámanankɛ : 1.0	Bambara
2.25	súurun : 1.0 , súuru : 1.0 , sùruntu : 2.0 , súruntu : 1.0 , sùrundu : 1.0	Verser
2.02	cíyɛn : 9.0 , tìɲɛ : 11.0 , tíɲɛ : 2.0 , ciyén : 2.0 , cìyɛn : 11.0	Vérité
2.00	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ : 1.0	Regard
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Approcher
1.99	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ : ? 1.0	?
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Salir
1.99	lé : 1.0 , lè : 4.0 , l̀è : 1.0 , le : 7.0 , dè : 4.0	(clan)
1.95	tàamasyɛnw : 2.0 , tàamashyɛn : 2.0 , táamashyɛnw : 1.0 , tàamashyɛnw : 2.0	Indiquer

Plan

Conclusions

- Corpus du bambara
  - Initiative déjà ancienne
  - Corpus plutôt TAL-compatible
- Analyse en parties du discours
  - Assez bonnes performances
  - Problématique de tonalisation
- Tonalisation
  - Utile à la morpho-syntaxe
  - Probablement utile pour les lemmes et gloses

Perspectives

- Travaux en cours
  - Coupler tonalisation et morpho-syntaxe
  - Essais de lemmatisation (jointe ?)
  - Statistiques et expériences avec les gloses
- Entités nommées (2 ème année)
  - Annotation en entités nommées
  - Translittération / transcription d'entités nommées
  - Expressions composées
- Malinké (3 ème année)
  - Travail sur un autre corpus
  - Proximité avec le bambara